Supporting Information

Using the classical twin model to decompose phenotypic variance into genetic and environmental components

In the classical twin design the observed resemblance in identical (monozygotic; MZ) twins versus nonidentical (dizygotic, DZ) twins allows the decomposition of the phenotypic variation into that due to genetic, shared environmental, and residual influences. Additive genetic variation (A) results from the sum of allelic effects across multiple genes. Shared environmental variation (C) results from environmental influences shared by family members, such as prenatal environment, home environment, and socioeconomic status. Residual variation (E) results from influences that are not shared by family members, such as idiosyncratic experiences, stochastic biological effects, as well as measurement error.

These different variance components can be estimated using twin data because MZ twins share all their genes, while DZ twins share on average half of their segregating genes. Any phenotypic correlation between MZ twins can be attributed to their shared genetic makeup and their shared environmental influences. As such, the observed correlation between MZ twins (r_{MZ}) equals A + C. The resemblance between DZ twins is due also to their shared environmental as well as their shared genetic influences - however, DZ twins share only 50% of their segregating genes on average. As such, the observed correlation between DZ twins (r_{DZ}) equals 0.5*A + C. The variance of each twin is due to their genetic make-up, shared environmental influences, and residual factors unique to each twin, i.e. A + C + E. These equations can be equivalently represented in a path diagram (Figure S1) or in algebraic form (Table S1).

Structural equation modelling (SEM) enables the estimation of A, C, and E while accounting for the effect of covariates (e.g. age), uneven MZ/DZ sample sizes, and missing

data, as well as determining the fit of the model to the data and confidence intervals around the parameter estimates.

Table S1. Variance/covariance matrix of MZ and DZ twins, with the variance for each twin on the diagonal (shaded dark grey) and the covariance between twins on the off-diagonal (shaded light grey).

MZ	*Twin 1*	*Twin 2*
Twin 1	$a^2+c^2+e^2$	a^2+c^2
Twin 2	a^2+c^2	$a^2+c^2+e^2$

DZ	*Twin 1*	*Twin 2*
Twin 1	$a^2+c^2+e^2$	$0.5a^2+c^2$
Twin 2	$0.5a^2+c^2$	$a^2+c^2+e^2$

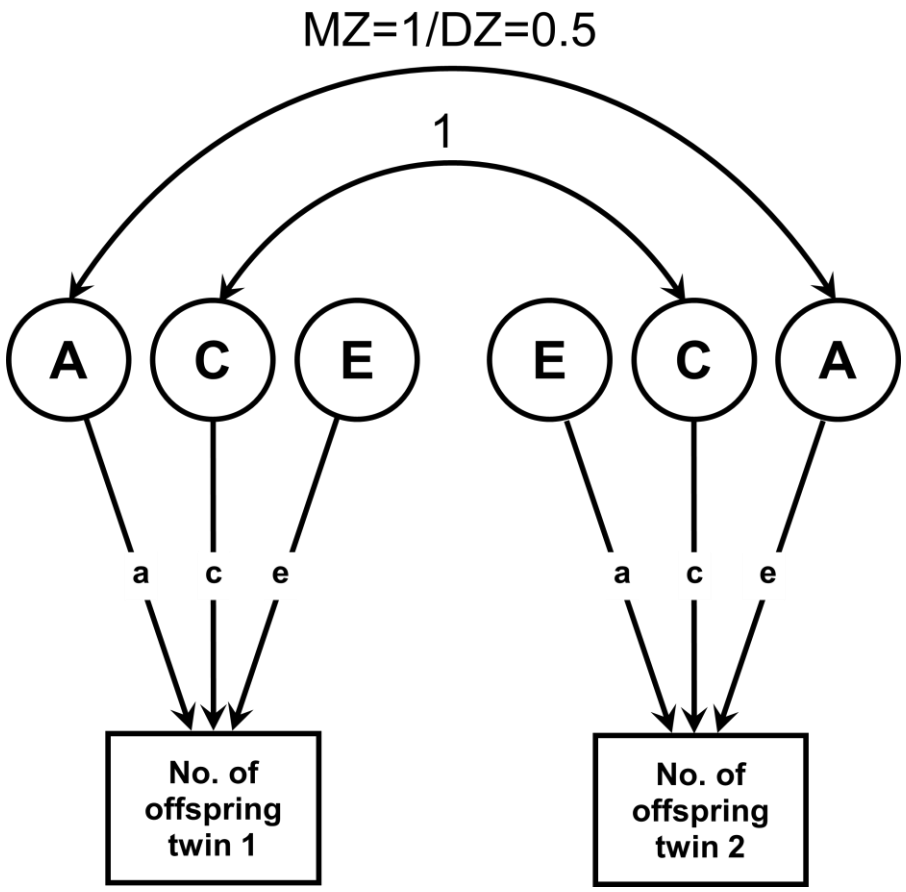

Figure S1. Graphical representation of the classical twin design. The boxes represent the observed variables (for twin 1 and twin 2 of a pair), and the circles represent the latent variables (A, C, and E), that influence the observed variables. The double-headed arrows show the correlations between the corresponding latent factors. A=genetic influences, C=shared environmental influences, E=residual influences.

Using the classical twin design to decompose the covariance between two traits

The univariate twin design described above can be extended to multiple traits. In a bivariate twin design, we use cross-twin cross-trait correlations (e.g. in this case the number of offspring in one twin and the number of grandoffspring in the other twin) to partition the covariation between the traits into genetic and environmental sources in the same way as we do for variation in a single trait. If the cross-twin cross-trait correlation is higher in MZ twins than in DZ twins it indicates a genetic contribution to the covariation between the traits.

More specifically, the bivariate design allows us to investigate the extent to which overlapping genetic factors (genetic correlation, r_A or r_g), shared environmental factors (shared environmental correlation, r_C) or residual factors (residual correlation, r_e) underlie the observed correlations between traits. Figure S2 (a reproduction of Figure 1 in the main paper) presents a path diagram of the bivariate twin design (called a Cholesky decomposition), where two traits are measured in twin 1 and twin 2. The first latent genetic variable (A1) explains the genetic influences on the first trait (in this case, number of offspring) and the correlated genetic influences on the second phenotype (number of grandoffspring). The second latent genetic variable (A2) is uncorrelated with A1 and explains the remaining heritability of number of grandoffspring. Corresponding latent structures are estimated for C and E. Again, the variance/covariance structure represented in the path diagram can also be written algebraically, shown in Table S2.

Table S2. Schematic representation of the bivariate variance/covariance matrix. Trait variances are on the diagonal (light blue), within-individual-cross-trait covariances are shown in light green, within-trait-cross-twin covariances (i.e. twin pair covariances) in pink, and cross-twin-cross-trait covariances in grey. Trait 1 = number of offspring; Trait 2 = number of grandoffspring; (1/0.5) applies to MZ (1) and DZ (0.5) twins, respectively. The values of the twin correlations can be found in Table 2 of the main text.

		Twin 1		*Twin 2*	
		Trait 1	Trait 2	Trait 1	Trait 2
Twin 1	Trait 1	$a_{11}^2+c_{11}^2+e_{11}^2$	$a_{11}*a_{21}+ c_{11}*c_{21}+ e_{11}*e_{21}+$	$a_{11}*(1/0.5)*a_{11}+ c_{11}^2$	$a_{21}*(1/0.5)*a_{11} +c_{21}*c_{11}$
	Trait 2	$a_{11}*a_{21}+ c_{11}*c_{21}+ e_{11}*e_{21}+$	$a_{21}^2+a_{22}^2+ c_{21}^2+c_{22}^2 + e_{21}^2+e_{22}^2$	$a_{11}*(1/0.5)*a_{21}+ c_{11}*c_{21}$	$a_{22}*(1/0.5)*a_{22} +c_{22}^2$
Twin 2	Trait 1	$a_{11}*(1/0.5)* a_{11}+c_{11}^2$	$a_{21}*(1/0.5)*a_{11} +c_{21}*c_{11}$	$a_{11}^2+c_{11}^2+e_{11}^2$	$a_{11}*a_{21}+ c_{11}*c_{21}+ e_{11}*e_{21}+$
	Trait 2	$a_{11}*(1/0.5)*a_{21}+ c_{11}*c_{21}$	$a_{22}*(1/0.5)*a_{22} +c_{22}^2$	$a_{11}*a_{21}+ c_{11}*c_{21}+ e_{11}*e_{21}+$	$a_{21}^2+a_{22}^2+ c_{21}^2+c_{22}^2+ e_{21}^2+e_{22}^2$

Following the tracing rules of path analysis (1), the parameters in Figure S2 and Table S2 can be converted into familiar statistics such as the heritability of each trait and the genetic correlation between them, along with the corresponding statistics for C and E influences. The formulas are presented in Box 1 below, and further details can be found in Loehlin (2).

The genetic correlation is easy to misunderstand. Analogous to calculation of the normal (phenotypic) correlation coefficient, the genetic correlation is calculated by dividing the genetic covariance between two traits by the square root of the product of the genetic variances of the two traits. In Figure S2, we can see that the genetic covariance can be traced through paths a11 and a21 (i.e. a11*a21); the genetic variance of trait 1 is up and back a11

(i.e. a11*a11); and the genetic variance of trait 2 is up and back a21 plus up and back a22 (i.e. a21*a21 + a22*a22). Thus the formula for the genetic correlation is $\frac{a11*a21}{\sqrt{a11^2}*\sqrt{a21^2+a22^2}}$ which reduces to the relevant formulas given in Box 1.

Genetic correlation measures the extent to which the genetic variance of each trait overlaps, with a genetic correlation of zero reflecting no overlap in the genetic variance and 1 reflecting complete overlap (i.e. the same genetic factors account for variation in the two traits). A genetic correlation of 1 does not necessarily imply a strong phenotypic correlation, because the latter is also a function of the heritabilities of the traits. Nor does a genetic correlation of 1 imply that the shared genetic influences are of the same magnitude in both traits. Further, a genetic correlation of 1 does not illuminate causation – for example, it could reflect a situation where the same pleiotropic genetic factors directly cause variation in trait A and trait B, or it could reflect a situation where genetic factors cause variation in trait A which in turn causes variation in trait B (and trait B has no independent genetic factors directly influencing its variance). The latter situation is likely to be the case for the relationship between offspring and grandoffspring.

Genetic correlations of 1 have been observed between closely related traits (e.g. between Major Depressive Disorder and Generalized Anxiety Disorder in females (3, 4)) and between the same trait measured at different time-points (e.g. drug use measured during adolescence and then again approximately 5 years later (5)).

Box 1. Formulas (referencing the parameters in Figure S2 and Table 3) for variance components of number of offspring and grandoffspring and for their genetic, shared environmental, and residual correlations. Variances are assumed to be standardised.

Variance in number of offspring due to A, C, and E:

A_{offs} = a11* a11

C_{offs} = c11* c11

E_{offs} = e11* e11

Variance in number of grandoffspring that is due to A, C, and E:

$A_{grandoffs}$ = a21 * a21 + a22 * a22

$C_{grandoffs}$ = c21 * c21 + c22 * c22

$E_{grandoffs}$ = e21 * e21 + e22 * e22

Genetic, shared-environmental, and residual correlations between number of offspring and grandoffspring:

$$r_A = \frac{a11*a21}{\sqrt{a11^2}*\sqrt{a21^2+a22^2}} = \frac{a21}{\sqrt{Agrandoffs}}$$

$$r_C = \frac{c21}{\sqrt{Cgrandoffs}}$$

$$r_E = \frac{e21}{\sqrt{Egrandoffs}}$$

Phenotypic correlation:

r_P = a11*a21 + c11*c21 + e11*e21

Proportion of phenotypic correlation due to genetic (1), shared environmental (2), and residual (3) factors:

(1) a11*a21/ r_P

(2) c11*c21/ r_P

(3) e11*e21/ r_P

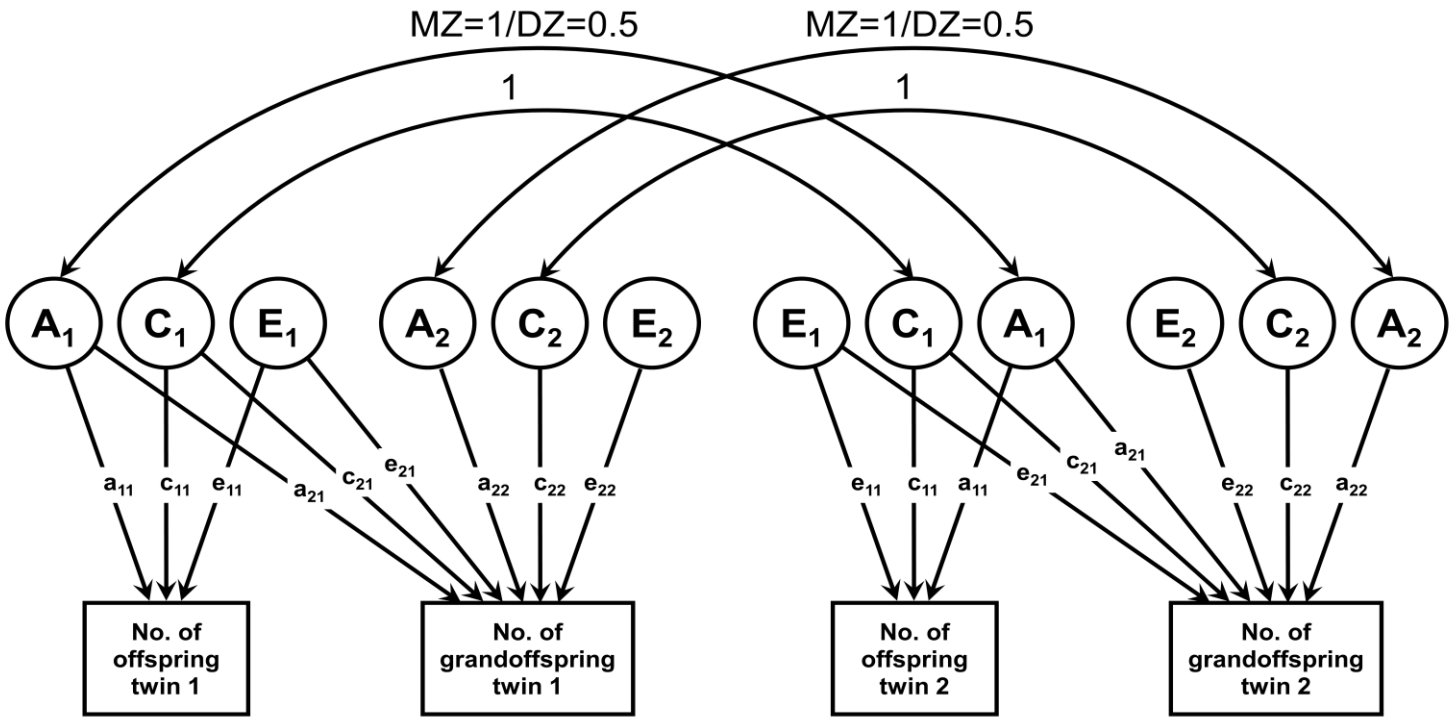

Figure S2. Path diagram of a Cholesky decomposition. The boxes represent the observed variables, and the circles represent the latent variables (A, C, and E) that influence the observed variables via the paths (arrows). The first latent genetic variable (A1) explains the genetic influences on number of offspring and the correlated genetic influences on number of grandoffspring. The second latent genetic variable (A2) is uncorrelated with A1 and explains the remaining heritability of number of grandoffspring. Corresponding latent structures apply to shared environmental (C) and residual (E) latent variables. Parameter estimates are reported in Table 3 in the main text.

Assessing the impact of truncation

To assess the impact of truncation on our analyzed twin sample (see main text Methods), we examine population cohorts not subject to truncation. We determine the proportion of the these cohorts having new offspring/ grandchildren at the affected ages (Tables S3 and S4), then use this information and the age distribution of our sample (Table S5) to estimate the proportion of our sample that have missing offspring/grandoffspring due to truncation.

In the case of number of offspring, we obtain the percentage of individuals in the full population cohort born 1933 who have offspring at each integer age prior to age 18 (i.e. the age at which the earliest-born individuals in our sample begin having their new offspring included in our data), shown in Table 3. We then multiply each percentage by the relevant percentage of our sample that would be subject to truncation if they had offspring at that age. For example, individuals born in 1915 (5.58% of our sample) would be affected by truncation if they had offspring before age 18 (~1.70% of the population, based on the 1933 rates) – .0558*.0170 – individuals born in 1916 (5.58%) would be affected by truncation if they had children before age 17 (~.45%) – .0558*.0045 – and so on. Summing all of these products yields an estimate that that only around 0.12% of the individuals in our sample are missing offspring due to truncation.

Males: (5.86*0.28 + 5.16*0.05 + 5.41*0.01)*.01 = 0.02%

Females: (5.34*3.18 + 5.94*0.86 + 6.00*0.09 + 5.59*0.01)*.01 = 0.23%

Total: (5.58*1.70 + 5.58*0.45 + 5.73*0.05 + 5.61*0.01)*.01 = 0.12%

For number of grandoffspring, we perform a similar analysis. We obtain the percentage of individuals in the population cohort born 1915 who have offspring at each integer age after 80 (i.e. the age at which the latest-born individuals in our sample stop having grandchildren counted), shown in Table 4. We then multiply each percentage by the relevant percentage of

our sample that would be subject to truncation if they had new grandoffspring at that age. For example, individuals born in 1929 (7.73%) would be affected by truncation if they had new grandoffspring after age 80 (~3.56%) – .0773*.0356 – individuals born in 1928 (7.71%) would be affected by truncation if they had new grandoffspring after age 81 (~2.89%) – and so on. Summing all of these products yields an estimate that that only around 1.3% of the individuals in our sample are missing offspring due to truncation.

Males: (5.86*0.03 + 5.16*0.18 + 5.41*0.32 + 5.62*0.48 + 5.81*0.66 + 7.42*0.84 + 6.89*1.08 + 7.27*1.34 + 6.68*1.64 + 6.27*2.03 + 6.03*2.48 + 7.75*2.98 + 7.84*3.61 + 8.07*4.27 + 7.90*5.13)*.01 = 1.98%

Females: (5.34*0.02 + 5.94*0.05 + 6.00*0.09 + 5.59*0.12 + 5.35*0.17 + 8.17*0.22 + 7.48*0.30 + 6.75*0.38 + 6.58*0.46 + 6.42*0.57 + 6.08*0.73 + 7.97*0.94 + 7.34*1.16 + 7.39*1.51 + 7.59*1.98)*.01 = 0.62%

Total: (5.58*0.02 + 5.58*0.11 + 5.73*0.20 + 5.61*0.30 + 5.57*0.42 + 7.82*0.53 + 7.20*0.69 + 6.99*0.87 + 6.63*1.05 + 6.36*1.30 + 6.06*1.61 + 7.87*1.96 + 7.57*2.39 + 7.71*2.89 + 7.73*3.56)*.01 = 1.29%

Note that the percentage of offspring/grandoffspring missing will be considerably lower than these values, because even individuals who are missing one or more offspring/grandoffspring will usually have other offspring/grandoffspring that are not missing.

Table S3. Percentage of the full population cohort born 1933 who have one or more offspring before each integer age whereby their data would be affected by truncation in our twin sample (i.e. <18 years of age).

	Percent of full population cohort born 1933 having offspring before 'Age'		
Age	Male	Female	Total
18	0.28	3.18	1.70
17	0.05	0.86	0.45
16	0.01	0.09	0.05
15	0.00	0.01	0.01
14	0.00	0.00	0.00
0-13	0.00	0.00	0.00

Table S4. Percentage of the population cohort born 1915 who have one or more grandoffspring after each integer age whereby their data would be affected by truncation in our twin sample (i.e. >80 years of age).

	Percent of population cohort born 1915 having new grandoffspring after 'Age'		
Age	Male	Female	Total
94	0.03	0.02	0.02
93	0.18	0.05	0.11
92	0.32	0.09	0.20
91	0.48	0.12	0.30
90	0.66	0.17	0.42
89	0.84	0.22	0.53
88	1.08	0.30	0.69
87	1.34	0.38	0.87
86	1.64	0.46	1.05
85	2.03	0.57	1.30
84	2.48	0.73	1.61
83	2.98	0.94	1.96
82	3.61	1.16	2.39
81	4.27	1.51	2.89
80	5.13	1.98	3.56

Table S5. The percentage of individuals in our twin sample born in each year.

	Percentage of twin sample		
Birth year	Male	Female	Overall
1915	5.86	5.34	5.58
1916	5.16	5.94	5.58
1917	5.41	6.00	5.73
1918	5.62	5.59	5.61
1919	5.81	5.35	5.57
1920	7.42	8.17	7.82
1921	6.89	7.48	7.20
1922	7.27	6.75	6.99
1923	6.68	6.58	6.63
1924	6.27	6.42	6.36
1925	6.03	6.08	6.06
1926	7.75	7.97	7.87
1927	7.84	7.34	7.57
1928	8.07	7.39	7.71
1929	7.90	7.59	7.73

SI Appendix References

1. Wright S (1934) The method of path coefficients. *Annals of Mathematical Statistics* 5:161-215.
2. Loehlin JC (1996) The Cholesky approach: A cautionary note. *Behavior Genetics* 26(1):65-69.
3. Kendler KS, Neale MC, Kessler RC, Heath AC, & Eaves LJ (1992) Major depression and generalized anxiety disorder: Same genes, (partly) different environments. *Archives of General Psychiatry* 49(9):716-722.
4. Kendler KS, Gardner CO, Gatz M, & Pedersen NL (2007) The sources of co-morbidity between major depression and generalized anxiety disorder in a Swedish national twin sample. *Psychological Medicine* 37(3):453-462.
5. Palmer RHC*, et al.* (2013) Stability and change of genetic and environmental effects on the common liability to alcohol, tobacco, and cannabis DSM-IV dependence symptoms. *Behavior Genetics* 43(5):374-385.